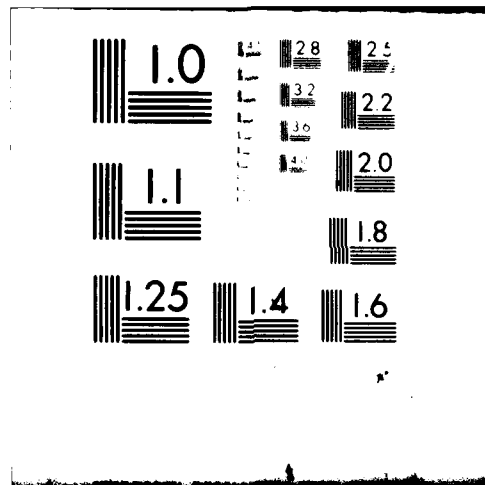END
DATE
FILMED
12-81
DTIC

LEVEL

NAVAL POSTGRADUATE SCHOOL
Monterey, California

DTIC
ELECTE
NOV 18 1981
S D
D

EFFECT OF OPERATOR MENTAL LOADING
ON VOICE RECOGNITION SYSTEM PERFORMANCE

by

J. W. Armstrong
and
Gary K. Poock

August 1981

AD A107477
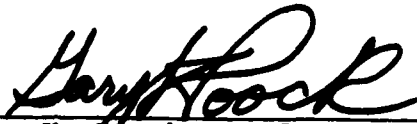
DTIC FILE COPY

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral J. J. Ekelund
Superintendent

D.A. Schrady
Acting Provost

Gary K. Poock, Professor
Department of Operations Research

Reviewed by:

Kneale T. Marshall, Chairman
Department of Operations Research

William M. Tolles
Dean of Research

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55-81-016	2. GOVT ACCESSION NO. AD A107477	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Effect of Operator Mental Loading on Voice Recognition System Performance.		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Gary K. Poock J. W. Armstrong		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS N0003980WR09041
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		12. REPORT DATE Aug 81
		13. NUMBER OF PAGES 60
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Naval Electronic Systems Command Washington, DC 20360		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

VTAG, Voice Recognition, Automatic Word Recognition, Mental Loading

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes an experiment in which military officers used voice recognition equipment under various mental loading conditions. Results showed voice recognition performance degraded as mental loading was imposed on the subjects.

DD FORM 1473 EDITION OF 1 NOV 65 IS OBSOLETE
1 JAN 73
S/N 0102-014-6601

251450

FOREWORD
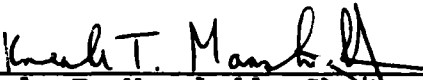
This investigation was sponsored by Mr. Frank Deckelman, NAVELEX, Code 330. The work was performed by the authors at the Naval Postgraduate School, Monterey, California.

This report is one in a series concerned with the possible applications of using voice recognition technology in command and control tasks.

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	

By
Distribution/

Availability Codes	
Dist	Avail and/or Special
A	-)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EFFECT OF OPERATOR MENTAL
LOADING ON VOICE RECOGNITION SYSTEM PERFORMANCE

A. OBJECTIVE AND BACKGROUND

The objective of this experiment was to determine if operator mental workload affected the performance of a voice recognition system comprised of a human operator and a discrete utterance voice recognition device. Specifically, the question addressed was: Would increased operator mental workload (with respect to that experienced during training of the recognition device) result in changes in his speech which would in turn result in degraded performance of the voice recognition system? A special vocabulary was used to ensure a baseline error rate with which to compare various mental loading levels. As such, it was expected that absolute error rates would be higher than those normally realized in real world operations. This experiment with mental loading has an integral relationship to previous motor loading research by Armstrong (1980).

B. SUBJECTS

Twenty-four subjects participated on a volunteer basis with no monetary or other incentive. Twenty-two of the subjects were students at the Naval Postgraduate School (NPS) and two were military staff members at NPS. They included 22 male military officers representing the United States Navy, Army, Air Force, Marine Corps and Coast Guard: one female civilian from the United States National Security Agency; and one male military

officer of the Canadian Forces. All subjects were between the
ages of 27 and 43 inclusive and the ranks of the military officers
ranged from Lieutenant to Commander and from Captain to
Lieutenant-Colonel inclusive.

Sixteen of the subjects, designated "little experience",
were subjects in a previous experiment by Poock (1980) and had
between two and ten hours experience on the voice recognition
system used in the experiment: - mean 6.2 hours; eight, designated
"no experience", had no experience on this equipment. Only two
of the subjects had experience - one half hour each - on the
Response Analysis Tester which was used to simulate operator
mental loading.

C. EQUIPMENT USED

1. Response Analysis Tester (RATER)

The General Dynamics Response Analysis Tester (RATER, Model 3)
shown in figure 1 was used to simulate operator mental loading.
Brady (1968) described the Rater as a "psychomotor testing in-
strument designed to provide sensitive, reliable measurement of
any impairment of response speed/accuracy and short-term memory
for patterned or color stimuli." Long and Fishburne (1973)
provide normative RATER performance data for a student naval
aviator population and reference several studies in which the
RATER was used. Newsom, Brady and O'Laughlin's study (1966)
of performance in a revolving space station simulator found that
turning the head while in a rotating environment resulted in
degraded short term memory as measured on the RATER.

2
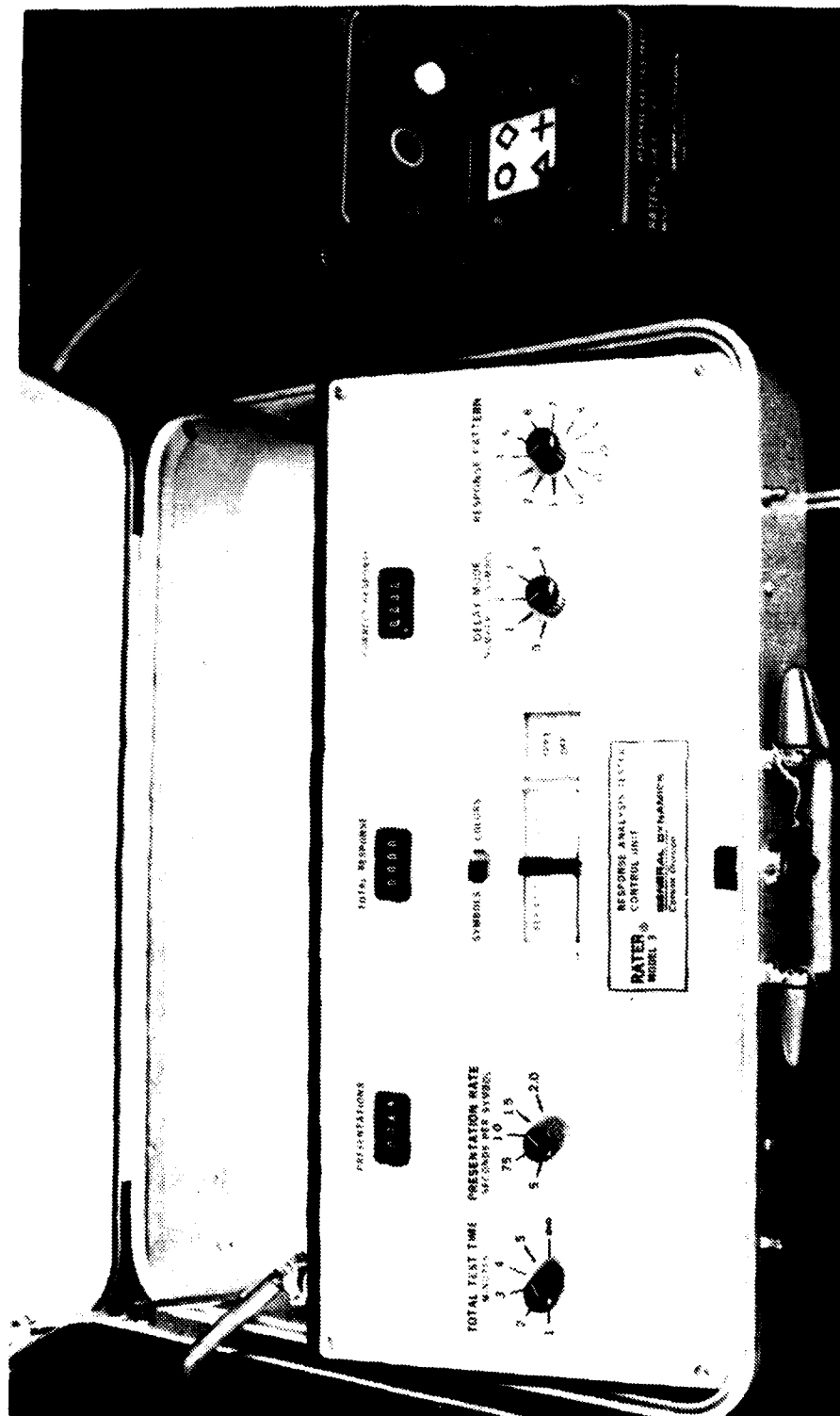
Subject console

Experimenter console

FIGURE 1. RESPONSE ANALYSIS TESTER (RATER)

The RATER consisted of a small subject console which contained a display window and four response buttons in a two by two arrangement and a larger experimenter console which contained the controls and digital counters. These counters were used in the derivation of subject RATER performance data.

The RATER was used to generate and display random sequences of four individual symbols - a triangle, a circle, a cross and a diamond - in the window of the subject console. Symbols were presented at a constant rate of one symbol every 1.5 seconds. A response button on the subject console was associated with each of the four symbols and labelled accordingly.

Three different RATER "delay" modes were used - delay zero, delay one and delay two. While the nth stimulus of the sequence, St(n), was being displayed and before St(n+1) replaced it, the subject was required to press the correct response button in order to score a correct response. In delay zero the correct response button was the one which corresponded to the symbol comprising St(n). In delay one the correct response button for the nth stimulus was the one which corresponded to the symbol comprising St(n-1); in delay two the correct response button for the nth stimulus was the one which corresponded to the symbol comprising St(n-2). In other words, in delay zero the subject responded with the symbol which correlated to the symbol being displayed. In delay one, the correct response was the symbol which had appeared the previous trial. In delay two, the correct response was the stimulus symbol which had been presented two trials earlier, i.e. the subject had to remember two back instead of one back (delay one) or none back (delay zero).

4

The RATER was used solely as a device to load the subjects mentally, i.e. to load the subjects through tasking which was primarily decision-making in nature. The choice of stimuli presentation rate and delay modes was based on experience gained during a pilot study, the findings of other researchers, especially Long and Fishburne (1973), and the expected lack of RATER experience of the subjects.

2. <u>Voice Recognition System and Choice of Vocabulary</u>

A Threshold Technology Inc. Model T600 discrete utterance voice recognition system (which will hereafter be referred to as the T600) was used as the equipment component of the combined equipment plus human operator voice recognition system. The vocabulary used in this experiment consisted of 50 different utterances. Thirty were single words selected by the experimenter from the Listener's Answer Sheets of the Modified Rhyme Test, one of the four test types which have been commonly used in measuring intelligibility in speech communication (Kryter, 1972). Sixteen of these 30 words were eight pairs of rhyming words which, within each pair, differed only with respect to initial consonant - for example, "beat" and "peat". The other 14 words were seven pairs of non-rhyming but similar words which, within each pair, differed only with respect to final consonant - for example, "sap" and "sat". The other 20 utterances were chosen by the experimenter from single words commonly used in Command and Control environments; they were chosen to be more easily distinguished from each other and from the other 30 words of the vocabulary.

5

All words of the vocabulary were one or two syllables in length. Short words were deliberately selected to facilitate generation of as many T600 word recognition attempts as possible in the limited time that each volunteer subject was available. The vocabulary is listed by word type in Appendix A. A listing in the order in which the words were trained is attached to the written instructions initially given to subjects and is contained in Appendix C.

This particular vocabulary was chosen to increase the likelihood of recognition errors by the T600 for the following reason. (T600 recognition errors (RE's) are operationally defined in the Dependent Variables section.) Recognition accuracy with older Threshold Technology Inc. voice recognition equipment similar to the T600 and using more normal vocabularies (i.e. comprised entirely of more easily distinguished words) has often been better than 99%, as for example, in the studies by Martin and Grunza (1974), Scott (1975) and Scott (1978). This level of accuracy would produce an average of about one (or less) RE's per 100 spoken utterances. It was anticipated that if operator mental loading did affect recognition accuracy then the effect would be relatively small and, due to the discrete nature of RE's, would probably not be easily distinguishable if only one RE per 100 utterances were being observed - for example, a 20% increase in RE's would probably not be great enough to produce a sufficient number of increased RE observations to be statistically distinguishable from inherent random variation. However, if a vocabulary

6

could be chosen to produce approximately ten RE's per 100 utterances a 20% increase in RE's should be more easily distinguishable as this would result in an average observation of 12 RE's per hundred utterances.

An alternative method of detecting a small expected change in recognition accuracy would be to increase the number of utterances spoken by the subjects. This was not considered feasible here because of the greatly increased time which would be required of each of the volunteer subjects; the experimental design used required between 1.5 and two hours per subject. For this reason the former method, special vocabulary, was used.

3. Arrangement of Equipment Used

Figure 2 illustrates the functional relationships among the various experimental devices used in the experiment. A photograph of the experimenter control station is shown in figure 3. The subjects were seated one at a time in an Industrial Acoustics Co. Inc. Controlled Acoustic Environments booth. The subject console of the RATER was on a table in front of the subject.

A Maico Model MA-24B Dual Channel Research and Diagnostic Audiometer and headsets were used to provide oral communication between the subject and the experimenter. The experimenter could speak to the subject by depressing a "talk-over" switch. Another microphone, placed in the booth, was live at all times and permitted the experimenter to hear what was happening in the booth - in particular, what the subject said. A Sony model TC 124 cassette tape recorder was connected to permit simultaneous recording of the signals detected by the booth microphone and the signals that the subject received over his headset.
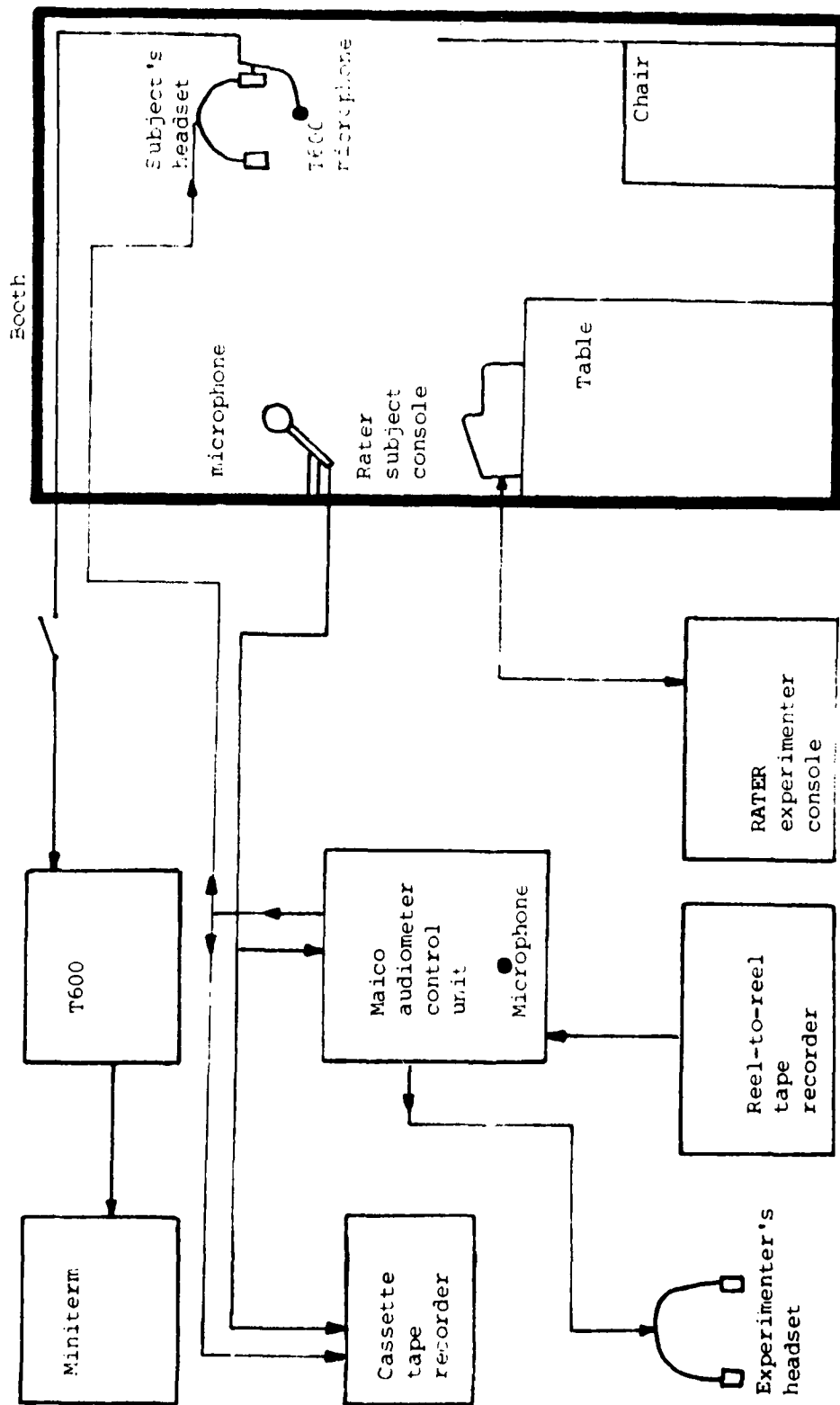
FIGURE 2. BLOCK DIAGRAM OF EXPERIMENTAL CONTROL SYSTEM

8
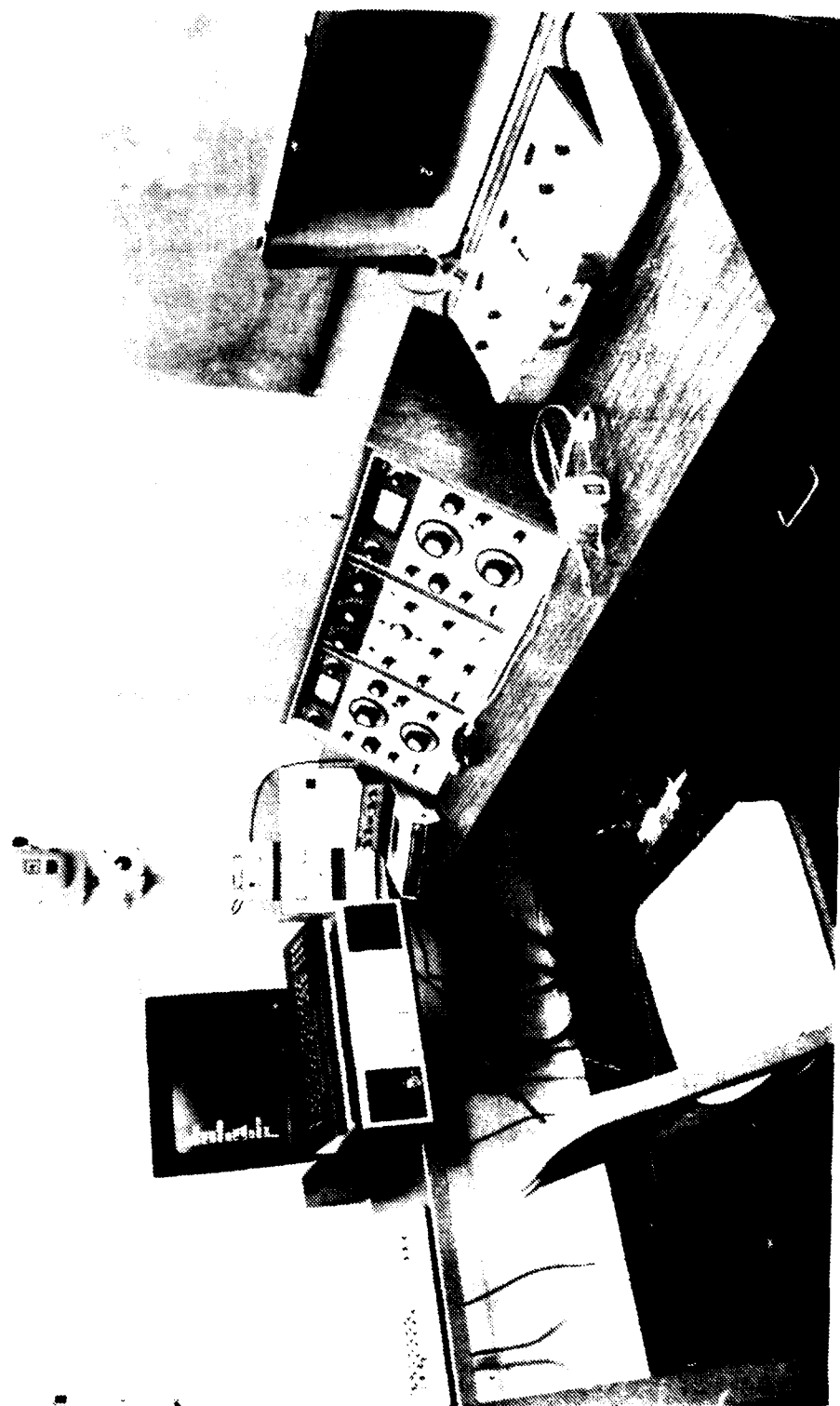
Miniterm   T600   Cassette   Maico   RATER
    tape recorder   Audiometer   (experimenter console)

FIGURE 3. EXPERIMENTER CONTROL STATION

The special T600 system noise-cancelling microphone was mounted on the subject's headset and connected only to the T600. The microphone ON/OFF switch was located outside of the booth.

A Computer Devices Inc. Model 1203 Miniterm portable terminal was connected to the T600 system in such a manner that when the T600 recognized an utterance the output string for that utterance was typed at the terminal. The T600 was programmed so that the ASCII output stream associated with each utterance of the vocabulary was simply the letters spelling the utterance followed by a carriage return and a line feed; thus, for example, if in the recognition mode the T600 "thought" that a subject said "attack", the word "attack" was displayed on the CRT on a separate line and printed at the terminal, also on a separate line. This provided the experimenter with a paper printout of T600 recognition activity which, with the correct utterances recorded on the cassette tape recorder, permitted thorough analysis of the data. Accurate, manual, real-time analysis by the experimenter using only the T600 CRT was infeasible primarily because of the rate at which the T600 was required to process signals for recognition - one word every three seconds.

An Akai model 4000DS Mk II reel-to-reel tape recorder was connected to the Maico Audiometer and used to present stimuli to the subject.

D. EXPERIMENTAL PROCEDURE

Subjects were tested one at a time during normal working hours. They were first required to complete the Subject Data

Form (Appendix B) and then read three pages of written instructions (Appendix C) which briefly introduced the experiment and provided general guidelines on inputting voice data to the T600. Remaining instructions to the subject were given orally by the experimenter.

"No experience" subjects only were next given a brief demonstration of the operation of the T600. For this stage, the T600 microphone and the headset on which it was mounted were removed from the booth and the microphone was reconnected outside of the booth so that the subject could immediately see what happened when speech signals were input to the T600. The importance of the guidelines which the subject had just read were demonstrated during this stage and the subject was allowed to familiarize himself with the T600 for about five minutes.

The T600 microphone and the headset on which it was mounted were then reconnected inside the booth. (The procedure from this point on pertains to all subjects.) The 50 word vocabulary was then trained one word at a time. The experimenter had all of the T600 controls outside of the booth and closely controlled the training process, requiring the subject to retrain words as necessary - for example, if a word was initially trained monotonously. The T600 was next put in the recognition mode and recognition of each word of the vocabulary was checked. Words which initially could not be recognized were retrained until they could be correctly recognized. If a word was correctly recognized immediately it was not checked further. Words not correctly recognized immediately were retrained if more than one

recognition error was obtained in three attempted recognitions of the word. Retrained words were rechecked and retrained again as necessary.

The subject next received, via his headset, a 2.5 minute tape recording of the 50 words of the vocabulary arranged in random order and presented at a constant rate of one word every three seconds. The subject was instructed to repeat the words one at a time for recognition by the T600. He was advised to try to repeat each word and to guess with a word in the vocabulary if he was uncertain.

Next the subject was briefed on the three RATER tasks that he would be performing - delay zero, delay one and delay two. He was advised that his RATER scoring would be number of correct responses minus number of incorrect responses, which included both omission and commission errors. The subject was also advised that he was not required to attain any particular proficiency levels on the RATER but that it was sufficient that he understood each of the tasks and did his best. He was then allowed to practice the three RATER tasks for up to 20 minutes. The RATER was used in the self-pace mode during parts of the practice if requested by the subject. In the self-pace mode the symbol displayed was replaced by the next symbol in the sequence only when a correct response was made.

When the subject advised the experimenter that he no longer wished to practice on the RATER the subject was given a combined 2.5 minute RATER delay one and word repetition for recognition

practice. The subject was played the same 2.5 minute tape re-
cording that he had heard earlier and was instructed as before
to repeat the words one at a time for recognition by the T600.
He was advised that this was the higher priority task but that
he was to simultaneously perform the RATER task was well as he
could with whatever capabilities he had remaining after attending
to the priority task. The subject was also reminded to be sure
to repeat each of the taped words and to guess with a word in
the vocabulary if he was uncertain.

The subject was then exposed to the four experimental con-
ditions corresponding to the four operator mental loading con-
ditions - no RATER task (NRT), RATER delay zero (RDO), RATER
delay one (RD1) and RATER delay two (RD2). These were designed
to create different levels of operator mental loading. Each con-
dition lasted five minutes and each of the 24 subjects received
the four conditions in a different order.

During condition NRT the subject was required only to repeat
two different consecutive random orderings of the words of the
vocabulary; these were presented to him over his headset as during
practice. The first time through the vocabulary in any condition
was referred to as the first half of the trial; the second time
was referred to as the second half of the trial. The first word
of the second half followed the last word of the first half with
the same spacing used within the two halves; the subject received
no cues that he was halfway through the trial. In each of the
conditions RDO, RD1 and RD2 the subject was similarly required to
repeat random orderings of the vocabulary (two different orderings

for each condition as in condition NRT); however, he was also required to perform simultaneously the appropriate RATER task. He was reminded that the repetition of words for recognition by the T600 was the higher priority task and to guess with a word from the vocabulary if he was uncertain, as during the combined practice. (The purpose of this instruction was to ensure that the T600 received the same, or at least nearly the same, utterances for recognition during each trial half and thus provide a common basis for comparison of T600 recognition errors.) By monitoring the T600 CRT display and RATER counters, listening to booth activity via the booth microphone, and post-experiment questioning of subjects, the experimenter ensured that subjects adhered to the instructions that they had been given.

Immediately after a subject completed each condition, and before he was allowed to leave the booth, he was instructed to complete the "Feeling Tone Checklist" shown in Appendix D in accordance with the instructions also shown in Appendix D. This checklist, developed by Pearson and Byars (1956), was administered to assess possible differential subjective fatigue after each of the four different mental loading conditions.

During the experimental conditions subjects <u>were not given feedback</u> on their RATER performance. During the practice sessions the only feedback given to subjects regarding T600 recognition of their speech was the knowledge of which words required re-training; <u>no feedback regarding T600 recognition performance</u>

14

was given to subjects during the experimental conditions. Those
subjects who indicated interest on their "Subject Data Sheets"
were individually briefed immediately after they completed
the last experimental condition concerning their RATER per-
formance, T600 recognition of their speech and the hypotheses
being tested.

Subjects were allowed to take short rest breaks as they
wished during the training and practice sessions and before each
of the four experimental conditions. A drinking fountain was
located nearby for any subjects who became thirsty or whose
throats became dry.

E. DEPENDENT VARIABLES

The following were calculated for each half of each trial:

1. T600 recognition errors (RE's)

2. Subject verbal errors.

In this experiment a T600 recognition error was operationally
defined to be a failure of the T600 to recognize correctly any
vocabulary word which a subject said; this included both incor-
rect recognition (for example, the subject said "beat" and the
T600 "thought" he said "peat") and rejection (for example, the
subject said "dip" and the T600 failed to recognize it and
emitted a "beep" sound). This definition is different from
most definitions of recognition error in the voice recognition
literature which do not include rejections - for example,
Martin and Grunza (1974). The operational definition used in
this experiment was considered more consistent with the aim of

15

this research - i.e. to answer the question: Would increased
operator mental workload (with respect to that experienced during
training of the recognition device) result in changes in his
speech which would in turn result in degraded performance of
the voice recognition system? It was believed that if the T600
rejected "dip" when said by a subject under condition RD2, but
not when said by the same subject under condition NRT, this
suggested changes in system performance as a result of changes
in the subject's speech and accordingly should be recorded and
analyzed.

A subject verbal error was defined as a failure of the
subject to repeat correctly the presented word. This failure
could be either a failure to respond (omission) or responding
with a non-vocabulary word or the wrong vocabulary word
(commission).

F. HYPOTHESES

The following hypotheses were to be tested.

1. Hypotheses Regarding T600 Performance

   a. H_o: The different levels of operator mental loading
            would not have different effects on T600
            recognition error rate.

   H_1: H_o false.
         It was expected that increased operator loading
         would result in increased recognition error
         rate (RER), i.e. RER(NRT) < RER(RDO) < RER(RD1)
         < RER(RD2)

16

b. H_O: The two trial halves would not have different effects on T600 recognition error rate.

H_1: H_O false.

c. H_O: "Little experience" subjects would generate the same T600 recognition error rate as "no experience" subjects.

H_1: H_O false.

It was expected that "little experience" subjects would generate a lower recognition error rate than "no experience" subjects.

2. Hypotheses Regarding Subject Performance

a. H_O: The different levels of operator mental loading would not have different effects on subject verbal error rate.

H_1: H_O false.

It was expected that increased operator loading would result in increased subject verbal error rate (VER), i.e. $VER(NRT) < VER(RDO) < VER(RD1) < VER(RD2)$ (This hypothesis was suggested by the research of Johnston (1975) who observed a significant detrimental effect of a simultaneous compensatory tracking task on speech intelligibility in noise.)

b. H_o: The two trial halves would not have different effects on subject verbal error rate.

   H_1: H_o false.

c. H_o: The different RATER delay modes used would not have different effects on subject RATER performance (score).

   H_1: H_o false.

   It was expected that subjects' RATER scores would decrease with increasing delay mode.

d. H_o: Subject subjective fatigue (as measured by the "Feeling Tone Checklist" of Pearson and Byars, 1956) would be the same for the four operator mental loading conditions.

   H_1: H_o false.

   It was expected that increased operator loading would result in increased subjective fatigue (SF), i.e. $SF(NRT) < SF(RDO) < SF(RD1) < SF(RD2)$

Subject T600 experience was not expected to affect subject verbal error rate or RATER performance and hypotheses regarding this were not devised. RATER performance was not recorded at the end of the first half of trials and hypotheses regarding RATER performance versus trial half were not devised.

G. **EXPERIMENTAL DESIGN**

A conceptual design for the experiment is shown in Figure 4. This is a three factor nested-factorial design. Each subject is nested within only one of the T600 experience level groups.

18
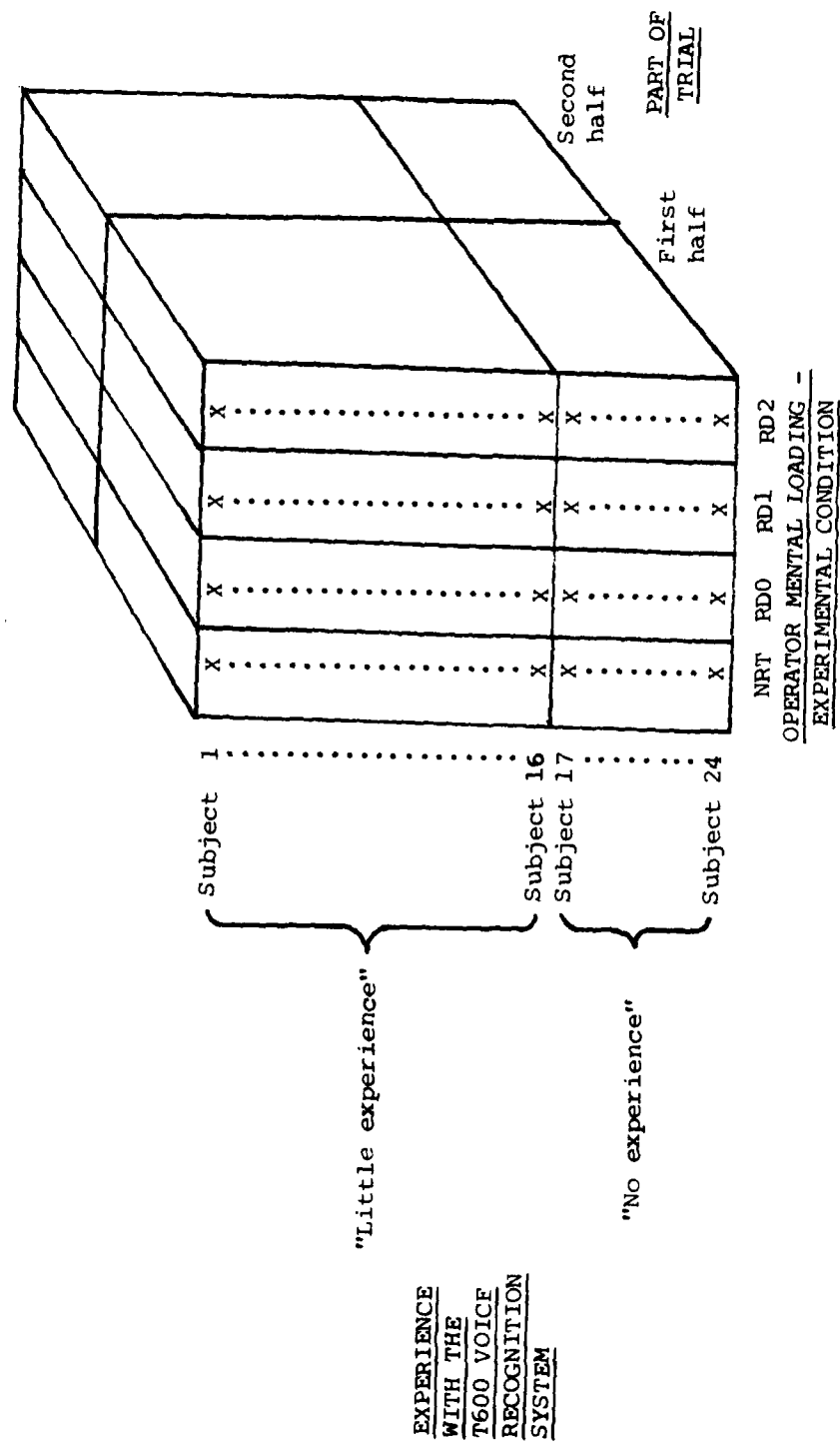
FIGURE 4. CONCEPTUAL DESIGN OF THE EXPERIMENT

Each of the 24 subjects was exposed to the four experimental conditions in a different order. Each condition was presented an equal number of times in each of the four order positions - first, second, third, and fourth - within both the "little experience" and "no experience" groups. Subject to these restrictions the order of presentation of the four conditions to any particular subject was assigned randomly.

Subject verbal error rate and T600 recognition error rate data were expected to be inherently Binomial in nature. In the case of subject verbal errors, the values of p, the probabilities of a subject verbal error, or equivalently, subject verbal error rates, were expected to be small. Because of this and because the values of n, number of words to be spoken, were relatively large, it was concluded that the distributions of subject verbal errors could be approximated by Poisson distributions and statistical methods based on the Poisson distribution were selected to test subject verbal error rate hypotheses.

In the case of T600 recognition error rates, the values of p, probabilities of a recognition error or recognition error rates, were expected to be too large to permit analyses based on the Poisson distribution. It was decided that a parametric analysis of variance would be used to test recognition error rate hypotheses; prior to this analysis the data would be transformed using the arcsin transformation, $y' = 2\arcsin(y^{1/2})$, to remove the relationship between the variance and mean expected because of the binomial nature of the data.

Non-parametric tests were selected for testing hypotheses regarding RATER scores and subjective fatigue because these data were not expected to meet the assumptions of parametric tests.

Because of the exploratory nature of this research, a level of significance, α, of .10 was elected during the design phase. This value was used in all tests of hypotheses.

H. RESULTS

1. Results for T600 Performance

Appendices E, F, G, H and I present separate confusion matrices for each of the four operator mental loading - experimental conditions (NRT, RD0, RD1 and RD2) and for all four conditions combined respectively. A matrix element a_{ij} of these matrices indicates the proportion of the time that the T600 "thought" that a subject said word j when the subject actually said word i. Mean T600 recognition error rates for each operator mental loading condition, trial half, subject T600 experience level and vocabulary word type, expressed in recognition errors per 100 spoken utterances, are shown in Table I. Results for the operational words show an error rate of 2.91% which is similar to the results of Poock (1980) and Armstrong (1980).

Figure 5 is a plot of the recognition error rate observations and Figure 6 a plot of the arcsin transformed recognition error rate observations. Figure 6 shows that the parametric analysis of variance homogeneity of variance assumption was adequately met. Since the parametric analysis of variance is quite robust regarding its Normality assumption (Scheffé, 1959), it was felt that this assumption also was adequately met and a parametric

21

TABLE I

MEAN T600 RECOGNITION ERROR RATES*

BY OPERATOR MENTAL LOADING - EXPERIMENTAL CONDITION

NRT	10.77%
RDO	13.18%
RD1	13.14%
RD2	13.60%

BY TRIAL HALF

First half	11.73%
Second half	13.61%

BY SUBJECT T600 EXPERIENCE LEVEL

"Little experience"	12.26%
"No experience"	13.50%

BY VOCABULARY WORD TYPE

Rhyming	25.17%
Non-rhyming but similar	12.33%
Operational	2.91%

OVERALL   12.67%

* Expressed in recognition errors per 100 spoken utterances.
  A recognition error was operationally defined in this research
  to be a failure of the T600 to recognize correctly any vocabulary
  word which S spoke and includes both incorrect recognition and
  rejection of vocabulary words; recognition errors do not include
  those cases where S spoke a word not in the vocabulary (or coughed,
  sighed, etc.) and the T600 generated a recognition.
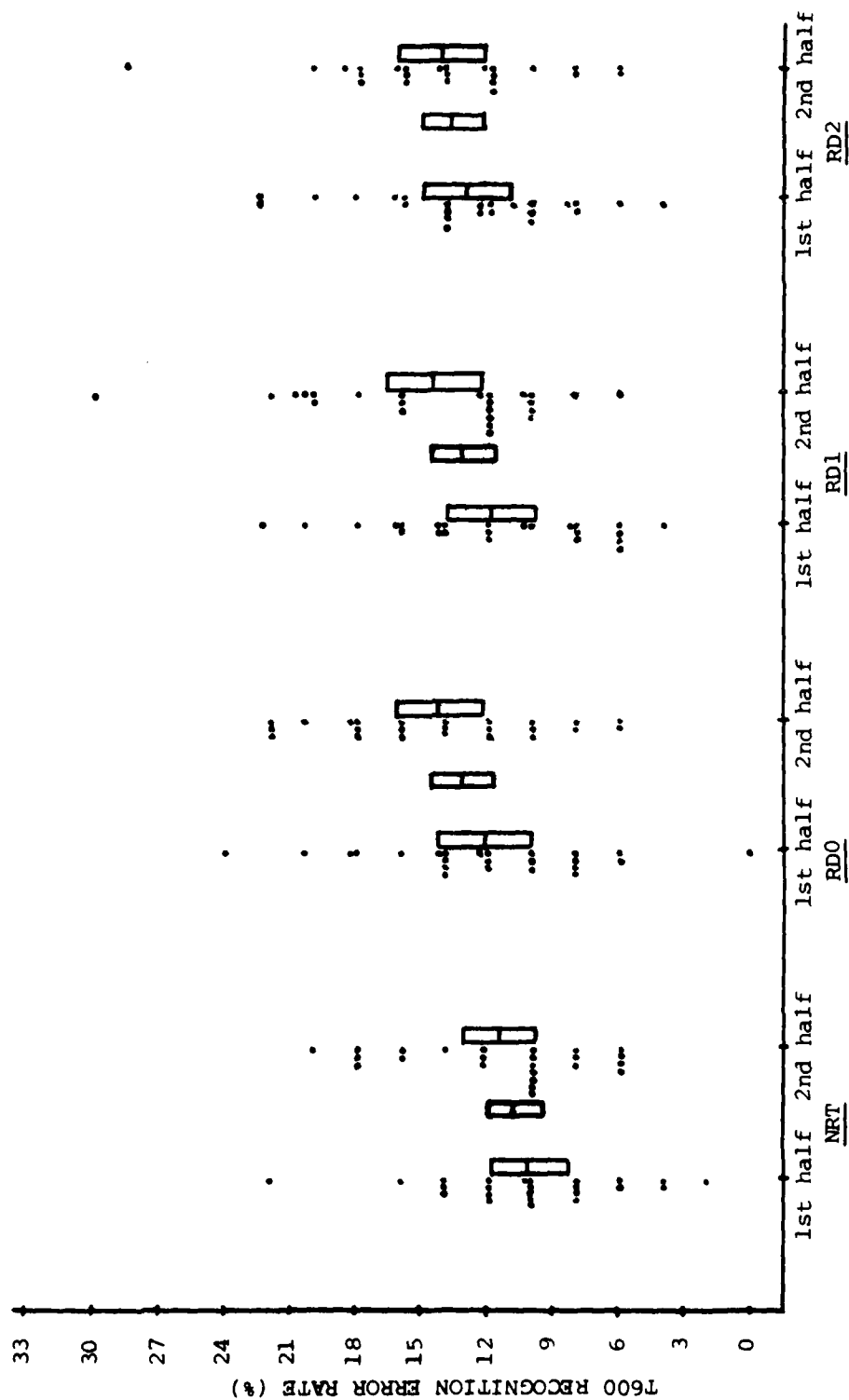
FIGURE 5. T600 RECOGNITION ERROR RATE OBSERVATIONS
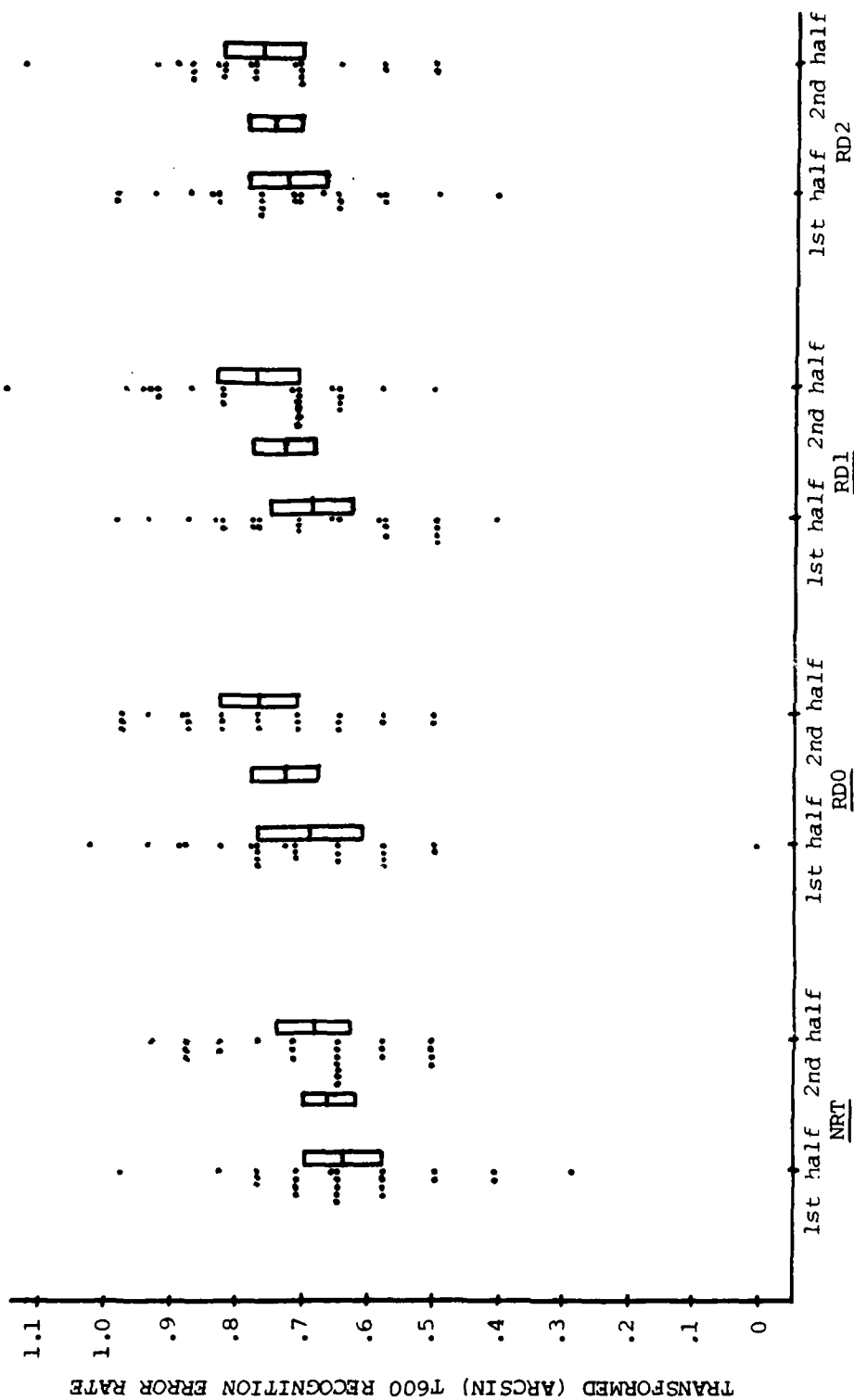
Three 95% confidence intervals are shown for each operator mental loading condition, one for the mean of the first half (left-most), one for the mean of the second half (right-most), and one for the overall mean (center). A level of significance, α, of .10 was selected in the experimental design.

23

Three 95% confidence intervals are shown for each operator mental loading condition, one for the mean of the first half (left-most), one for the mean of the second half (right-most), and one for the overall mean (center). A level of significance, α, of .10 was selected in the experimental design.

FIGURE 6. TRANSFORMED (ARCSIN) T600 RECOGNITION ERROR RATE OBSERVATIONS

analysis of variance (Winer, 1962) was performed on the arcsin transformed data. The results are summarized in Table II. The model for this analysis was:

$$Y_{ijkm} = u + L_i + H_j + E_k + S_{m(k)} + LH_{ij} + LE_{ik} + HE_{jk} + LHE_{ijk} + e_{ijm(k)}$$

where Y_{ijkm} = arcsin transformed recognition error rate for operator mental loading condition i, trial half j, T600 experience level k, and subject m; the range of Y_{ijkm} is 0 to π.

u = common experimental contribution to Y_{ijkm}

L_i = contribution of operator mental loading condition i, i = 1,2,3,4 (NRT, RD0, RD1, RD2)

H_j = contribution of trial half j, j = 1,2 (first half, second half)

E_k = contribution of T600 experience level k, k = 1,2 ("Little experience", "No experience")

$S_{m(k)}$ = contribution of subject m within T600 experience level k

m = 1,2, ..., 16 for k = 1

m = 1,2, ..., 8 for k = 2

$e_{ijm(k)}$ = random error

Subject effects were considered to be random; all others were considered to be fixed.

The analysis showed mental loading to be significant (F = 4.88, df = 3/66, p < .005). A parametric Range Test

TABLE II

ANALYSIS OF VARIANCE FOR T600 RECOGNITION ERROR RATE

Source	df	MS	F
Between Subjects	23		
E (T600 experience)	1	.0712	
Subj. w. groups	22	.0715	
Within Subjects	168		
L (Operator mental loading condition)	3	.0746	4.88*
E x L	3	.0114	
L x subj. w. groups	66	.0153	
H (trial half)	1	.1819	13.38*
E x H	1	.0001	
H x subj. w. groups	22	.0136	
L x H	3	.0058	
E x L x H	3	.0150	
L x H x subj. w. groups	66	.0201	

* $p < .005$

(Hicks, 1973) was performed to determine which operator mental loading conditions were statistically different (with respect to T600 recognition error rates) and it was found that the only significant differences (α = .10) were those between condition NRT and each of the other three conditions, RD0, RD1 and RD2. The analysis also showed recognition error rate to be higher in the second half of trials than in the first half (F = 13.38, df = 1/22, p < .005). Subject T600 experience level was not significant (F < 1). No interactions were significant (all F's < 1). Figure 7 shows recognition error rate versus operator mental loading condition for each trial half.

Subjects were instructed to repeat each vocabulary word heard and to guess with a word in the vocabulary if uncertain of the word. The purpose of this instruction was to ensure that the T600 received the same, or at least nearly the same, utterances for recognition during each trial half, i.e. each vocabulary word once, and thus provide a common basis for comparison of T600 recognition errors. Despite the instruction a total of 53 instances arose where subjects either did not speak any word or spoke a word not in the vocabulary; these are tabulated in Appendix J. T600 recognition errors, as operationally defined in this research, could not occur in these instances and the following adjustment was made to establish a reasonably common basis for comparison. If x T600 recognition errors occurred in a particular trial half for a subject and that subject made y errors of this type in the trial half, then the error rate observation on which the analysis was based was x/(50-y), not x/50.

27
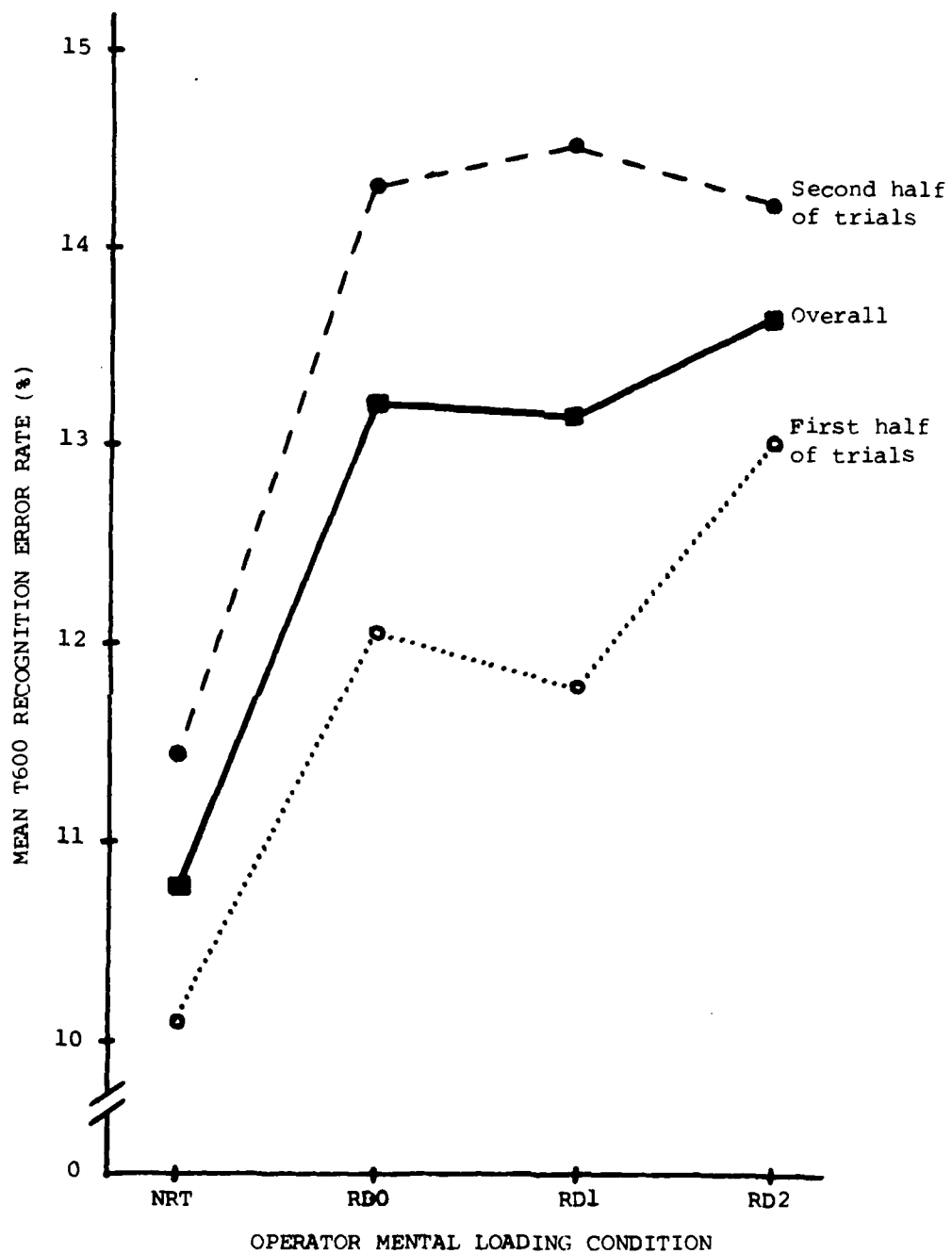
FIGURE 7. MEAN T600 RECOGNITION ERROR RATES

(in recognition errors per 100 spoken utterances)

28

2. Results for Subject Performance

Appendix J shows total subject verbal errors for each subject for each half of each trial under each operator mental loading condition. Mean subject verbal error rates for each mental loading condition, trial half, subject T600 experience level and vocabulary word type, expressed in subject verbal errors per 100 words presented to the subject for repetition (i.e. each word of the 50 word vocabulary twice), are shown in Table III.

Tests based on the Poisson distribution (Cox and Lewis, 1966) were performed on the subject verbal error rate data. It was concluded that the operator mental loading condition effect was significant ($p < .01$, $\alpha = .10$) and that the trial half effect was not significant ($p > .8$, two-tailed test, $\alpha = .10$).

Subject RATER scores are shown in Appendix K: A non-parametric Friedman two-way analysis of variance (Siegel, 1956) was performed on the RATER scores and it was concluded that scores varied by delay mode ($\chi_r^2 = 42.75$, df = 2, $p < .0005$, $\alpha = .10$). A non-parametric test proposed by Nemenyi (in Kirk, 1968, p. 497) was performed to determine which pairwise comparisons of RATER scores were significant; it was found that all pairwise differences were significant ($p < .05$) with RATER performance declining as the delay mode increased from 0 to 1 to 2.

The results of the subjective fatigue inquiry are shown in Appendix L. Numerical scores shown were obtained by multiplying the number of items scored "better than" by two and adding the number of items scored "same as", as recommended by Pearson and Byars (1956). A non-parametric Friedman two-way analysis of

TABLE III

MEAN SUBJECT VERBAL ERROR RATES*

BY OPERATOR MENTAL LOADING - EXPERIMENTAL CONDITION

NRT	.42%
RD0	.63%
RD1	1.04%
RD2	1.38%

BY TRIAL HALF

First half	.90%
Second half	.83%

BY SUBJECT T600 EXPERIENCE LEVEL

"Little experience"	.98%
"No experience"	.63%

BY VOCABULARY WORD TYPE

Rhyming	.81
Non-rhyming but similar	1.15%
Operational	.70%

OVERALL   .86%

*Expressed in subject verbal errors per 100 vocabulary words
presented to S via the headset. A subject verbal error was
defined in this research to be a failure of the subject
to repeat correctly the presented vocabulary word. This
failure could be either a failure to respond (omission) or
responding with a non-vocabulary word or the wrong vocabulary
word (commission).

variance was performed on this data and it was concluded that subjective fatigue was the same for the four operator mental loading conditions $(\chi_r^2 = 3.09, df = 3, p > .3, \alpha = .10)$.

The unexpected difference between the mean subject verbal error rates for "Little experience" and "No experience" subjects, shown in Table III, prompted the author to test whether or not this difference was significant. A test based on the Poisson distribution was performed and concluded that the difference was significant $(p < .10$, two-tailed test, $\alpha = .10)$.

3. General Results

The following were investigated graphically:

a. T600 recognition error rate versus subject verbal error rate; and,

b. RATER scores versus subject verbal error rates.

No relationships were apparent. Spearman rank correlation coefficients between subject RATER scores and T600 recognition error rates were calculated for each delay mode; none were found to be significant $(r_s(RD0) = -.110; r_s(RD1) = .127; r_s(RD2) = -.214; r_s(critical) = \pm.343$, two tailed test, $\alpha = .10)$.

I. DISCUSSION

Operator mental loading had a significant differential effect on subject verbal error rate, as expected, but trial half did not. "Little experience" subjects had a higher subject verbal error rate than "no experience" subjects; why this occurred is not known.

The subjective fatigue checklist used did not disclose significant differences between any of the four operator mental loading - experimental conditions. This was probably partly because the effects of order of presentation of the conditions dominated any possible condition effects during subjects scoring of the checklists. (Several subjects advised the experimenter after a RATER condition that the condition was more fatiguing than condition NRT but they had to score the RATER condition higher because it was the last, or next to last, condition and the subject felt good because the end was at hand.)

The following hypotheses were confirmed.

1. Operator mental loading affected the performance of the voice recognition system in that T600 recognition error rates in the three conditions involving concurrent RATER tasking were 23% greater than the error rate of the no RATER task condition.

2. Performance of the voice recognition system during the first 2.5 minutes of a trial differed from that during the second 2.5 minutes. A future experiment will investiate this possible degradation over time.

3. T600 recognition error rates were not statistically different for "no experience" and "little experience" (with respect to the T600) subjects. This may simply be due to the limited experience of even the most experienced subject who had only 12 hours previous experience.

It must be emphasized that the recognition error rates obtained with the T600 in this experiment are at least ten times what has commonly been found. These higher recognition error rates were deliberately sought by the experimenters (as discussed earlier) and are primarily due to the vocabulary selected. The average error rate on the 20 operational vocabulary words was 2.91%; the average error rate on the 30 words taken from the Modified Rhyme Test was 19.18%. A non-parametric Friedman two-way analysis of variance was performed and concluded that recognition error rate differed by vocabulary word type (rhyme, non-rhyme but similar, and operational) ($\chi_r^2 = 45.06$, df = 2, p < .0005). A non-parametric test proposed by Nemenyi (in Kirk, 1968, p. 497) was performed to determine which pairwise comparisons of recognition error rates were significant; it was concluded that all pairwise differences were significant (p < .01).

After the a priori hypotheses had been tested it was suggested that the T600 recognition error hypotheses be retested using only operational vocabulary word data. This was done using tests based on the Poisson distribution. The analysis showed the operator mental loading condition effect to be significant (p < .10), as it was when using the whole vocabulary. The trial half effect was found to be not significant (p > .2). It is not known whether this result indicates that the trial half difference observed when using the whole vocabulary was not present with the operational words or whether it indicates that the test using just the operational words was not powerful enough to detect the difference. This uncertainty will be investigated

33

in a future experiment. The analysis also showed that "no
experience" subjects generated higher recognition error rates
than "little experience" subjects (p < .10, two-tailed test,
α = .10).

This may be due to the fact that the "little experience"
subjects had more experience inputting the operational words
of the vocabulary than the "no experience" subjects. Most of
the operational words used were also used in the experiment
by Poock (1980) in which all of the "little experience" subjects
participated; none of the "no experience" subjects participated
in that experiment.

LIST OF REFERENCES

1. Armstrong, J. W. <u>The Effects of Concurrent Motor Tasking on Performance of a Voice Recognition System</u>, Naval Postgraduate School, Master's Thesis, September, 1980.

2. Brady, J. F., <u>Percuptual-motor Performance Degradation as a Function of Passively Induced Vestibular Coriolis Stimuli</u>, General Dynamics, Convair Division, Report GDC-ERR-1287, 1968.

3. Cox, D. R. and Lewis, P. A. W., <u>The Statistical Analysis of Series of Events</u>, Methuen and Co. Ltd., 1966.

4. Hicks, C. R., <u>Fundamental Concepts in the Design of Experiments</u>, Second Edition, Holt, Rinehart and Wiston, 1973.

5. Johnston, M. E., <u>The Effect of a Tracking Task on Speech Intelligibility in Noise</u>, Royal Aircraft Establishment, Technical Report 75014, (AD-A014 516) March 1975.

6. Kirk, R. E., <u>Experimental Design Procedures for the Behavioral Sciences</u>, Brooks/Cole Publishing Company, 1968.

7. Kryter, K. D., in Van Cott, H. P. and Kinkade, R. G. (editors), <u>Human Engineering Guide to Equipment Design</u>, Revised Edition, 1972.

8. Long, G. M. and Fishburne, R. P., <u>Performance Norms and Research Applications of the Response Analysis Tester (RATER)</u>, Naval Aerospace Medical Research Laboratory, Aerospace Psychology Technical Memorandum 73-1, 7 September 1973.

9. Martin, T. B. and Grunza, E. F., <u>Voice Control Demonstration System</u>, Air Force Avionics Laboratory, Report TR-74-174, (AD-B004 928L) March 1974.

10. Newsom, B. D., Brady, J. F. and O'Laughlin, T. W., <u>Study of Performance in a Revolving Space Station Simulator as a Function of Head Rotation about Y and Z Cranial Axes</u>, General Dynamics, Convair Division, Report GDC DBD 65-043-12, 21 November 1966.

11. Pearson, R. G. and Byars, G. E. Jr., <u>The Development and Validation of a Checklist for Measuring Subjective Fatigue</u>, Air University School of Aviation Medicine, USAF, Report 56-115, December 1956.

12. Poock, G. K., <u>Experiments with Voice Input for Command and Control</u>, Naval Postgraduate School, Technical Report NPS-55-80-016, April 1980.

13. Scheffé, H., <u>The Analysis of Variance</u>, John Wiley and Sons, Inc., 1959.

14. Scott, P. B., _Alpha-numeric Extraction Technique_, Rome Air Development Center, Report TR-75-287, (AD-B008 303L), November 1975.

15. Scott, P. B., _Word Recognition_, Rome Air Develoment Center, Report TR-78-209, (AD-A061 545) September 1978.

16. Siegel, S., _Nonparametric Statistics for the Behavioral Sciences_, McGraw-Hill, 1956.

17. Winer, B. J., _Statistical Principles in Experimental Design_, McGraw-Hill, 1962.

APPENDIX A

VOCABULARY LISTING (BY WORD TYPE)

RHYMING

gale	tale	gold	cold
game	came	bark	park
tip	dip	big	pig
beat	peat	ten	den

NON-RHYMING BUT SIMILAR

sap	sat	peas	peace
race	raze	save	safe
lake	late	kit	kid
mad	mat		

OPERATIONAL

list	course	attack	refuel
time	plot	bingo	cancel
speed	air	report	proceed
dive	fire	distance	label
drop	launch	copy	station

A vocabulary listing in the order in which the words were
trained is attached to the written instructions initially
given to subjects and shown in Appendix C.

APPENDIX B

SUBJECT DATA SHEET

Subject number: _____ Name: _____ Age: _____

Time/date: _____ Service: _____

Rank: _____ MOS (in words): _____

Do you object to being taperecorded during the experiment? If you do, stop filling out this form and advise the experimenter now; otherwise, continue.

How many hours experience have you had on voice recognition equipment in the last six months?

_____hours (approximately)

How many hours experience have you had on reaction measurement devices in the past year?

_____hours (approximately)

Do you have a speech or hearing impediment? Yes   No
                                             (circle one)

Do you want a post participation briefing on your performance and on the hypotheses being tested by the experimenter? Note that if you request such a briefing, you must agree not to discuss this with anyone other than the experimenter so that no subject will learn what results are expected prior to his participation in the experiment; such prior knowledge could invalidate the results of the experiment.

                                             Yes   No
                                             (circle one)

After you have completed participation in the experiment you will be asked to write below any comments which you think may be useful to the experimenter. If you have any questions now, please ask the experimenter. Otherwise, give him this form now and start reading the pages titled "INTRODUCTORY REMARKS/ RECOGNIZER VOCABULARY TRAINING".

_____

POST EXPERIMENT COMMENTS

(continue on reverse side if this space is insufficient)

THANK YOU FOR YOUR PARTICIPATION

WRITTEN INSTRUCTIONS

INTRODUCTORY REMARKS / RECOGNIZER VOCABULARY TRAINING

INTRODUCTORY REMARKS

This experiment involves analysis of a combined human operator / voice recognition equipment system under various conditions of operator mental loading. The actual experiment will be carried out in a sound-proof booth and subject - experimenter communication during the actual experiment will be via the booth intercom system; however, you may remove the headset assembly during break periods and leave the booth.

CAUTION: The mounting of the voice recognizer microphone on the headset assembly is very delicate, easily damaged, and difficult to repair. Please be careful while handling this assembly.

Please carry out the experiment exactly as directed and do not discuss your performance with anyone other than the experimenter as inappropriate subject prior knowledge could invalidate the results.

VOICE RECOGNIZER VOCABULARY TRAINING

The 50 word vocabulary being used with the voice recognizer in this experiment is attached to these instructions. You will be required to repeat each word of this vocabulary ten times to train the recognizer to recognize your particular vocalizations of each word. To facilitate recognition by the voice recognizer, you should include in the ten repetitions

as many as possible of the different ways you might say the word in normal speech; for example, use different intonations and emphasis, and small variations in volume.

In order to keep track of the number of times you say each word, and to reduce breath noise, it is best to speak the 10 repetitions in several groups. For example, if the word is zero, it is better to group them as:

	000-000-0000
or	000-000-000-0
rather than as	0000000000
or	0-0-0-0-0-0-0-0-0-0

Please observe the following guidelines while inputting voice data to the recognizer both during training and later during the actual experiment.

a. Speak each word crisply and quickly but do not over-pronounce; for example, words ending in "t" - delete final "t" if more natural.

b. Be sure to leave a distinct pause (specifically, at least one-tenth of a second of silence) between each word so that the recognizer can distinguish the end of one word from the beginning of the next. Similarly, do not leave a period of silence within a word or the recognizer will mistake it for two separate words.

c. Avoid breathing into the microphone at the end of words as this will generate false inputs to the recognizer.

d. Microphone location is very important and should be
     kept constant throughout the experiment; i.e., adjust
     it if it gets out of place. The experimenter will
     initially demonstrate correct microphone placement.

From this point on instructions will be given to you
verbally by the experimenter. Please advise him if you have
any questions now.

VOCABULARY LISTING (IN TRAINING ORDER)

0.	attack	25.	refuel
1.	list	26.	tip
2.	gale	27.	dip
3.	tale	28.	drop
4.	bingo	29.	lake
5.	sap	30.	late
6.	sat	31.	course
7.	time	32.	big
8.	gold	33.	pig
9.	cold	34.	report
10.	cancel	35.	kit
11.	peas	36.	kid
12.	peace	37.	plot
13.	speed	38.	beat
14.	game	39.	peat
15.	came	40.	proceed
16.	distance	41.	mad
17.	race	42.	mat
18.	raze	43.	fire
19.	copy	44.	ten
20.	bark	45.	den
21.	park	46.	label
22.	launch	47.	air
23.	save	48.	station
24.	safe	49.	dive

APPENDIX D

SUBJECTIVE FATIGUE CHECKLIST

Subject number _____ Experimental condition _____

FEELING TONE CHECK LIST

No.	Better than	Same as	Worse than	Statement
1.	( )	( )	( )	slightly tired
2.	( )	( )	( )	like I'm bursting with energy
3.	( )	( )	( )	extremely tired
4.	( )	( )	( )	quite fresh
5.	( )	( )	( )	slightly pooped
6.	( )	( )	( )	extremely peppy
7.	( )	( )	( )	somewhat fresh
8.	( )	( )	( )	petered out
9.	( )	( )	( )	very refreshed
10.	( )	( )	( )	ready to drop
11.	( )	( )	( )	fairly well pooped
12.	( )	( )	( )	very lively
13.	( )	( )	( )	very tired

Have you checked each statement?

INSTRUCTIONS FOR COMPLETING FEELING TONE CHECKLIST

People feel different at various times for various reasons. Some arise after a night's rest feeling "quite rested" while others may feel "a little tired". A hard day's work or a vigorous workout at the gym may make you feel "fairly well pooped"; yet, a shower, a cup of coffee, or merely a few minutes relaxing in a comfortable chair may make you feel "very refreshed".

I would like to find out how you feel right now. On the accompanying sheet, you will see 13 statements which describe different degrees of freshness or peppiness and tiredness. For each statement you will have to determine in your own mind whether you feel at this instant (1) "Better than", (2) the "Same as", or (3) "Worse than" the feeling described by that statement. Having done this you will then place an "X" in the appropriate box.

Consider the following example:

No.	Better than	Same as	Worse than	Statement
0.	( )	( )	( )	somewhat tired

If right now you felt "somewhat tired" you would place an "X" in the box marked "Same as". If, however, you felt fresh or full of pep you would check the box marked "Better than" because you would be feeling better than "somewhat tired". On the other hand, if you felt exhausted you would place an "X" in the box marked "Worse than".

Take each statement in order; do not skip around from one to another. Read each statement carefully so that you under-stand what it means. It may help you to understand some state-ments if you mentally insert the words "I feel" or "I am" before the statement.

This is not a test. You have all the time you need.

APPENDIX E

CONFUSION MATRIX FOR OPERATOR MENTAL LOADING – EXPERIMENTAL CONDITION NRT

The matrix element a_{ij} indicates the proportion of the time that the T600 "thought" that a subject said word j when the subject actually said word i.] The last column, annotated "beep", indicates the proportion of the time that the T600 rejected vocabulary word i and "beeped".

Several rows do not sum to exactly 1.00; this is because of inaccuracy in the elements of the row caused by rounding to two decimal places.

	GALE	TALE	GAME	CAME	TIP	DIP	BEAT	PEAT	GOLD	COLD	BANK	PARK	BIG	PIG	TEN	DEN	SAP	SAT	RACE	RAZE	LAKE	LATE	MAD	MAT	PEAS	PEACE	SAVE	SAFE	KIT	KID	LIST	TIME	SPEED	DIVE	DROP	COURSE	PLOT	AIR	FIRE	LAUNCH	ATTACK	BINGO	CANCEL	DISTANCE	COPY	REFUEL	REPORT	PROCEED	LABEL	STATION	"BEEP"
GALE	.98	.02																																																	
TALE	.19	.81																																																	
GAME			.79	.21																																															.02
CAME			.21	.79																																															.08
TIP					.63	.29																																													.02
DIP					.19	.69																																													.04
BEAT							.70	.23																																											
PEAT							.25	.69																																											
GOLD									.85	.15																																									
COLD									.08	.92																																									
BANK											.75	.19																																							
PARK											.23	.71																																							
BIG													.81	.17																																					.02
PIG													.25	.71																																					.02
TEN															.79	.19																																			
DEN															.17	.81																																			
SAP																	.81	.15																																	
SAT																	.04	.96																																	
RACE																			.90	.10																															
RAZE																			.02	.98																															
LAKE																					.87	.13																													
LATE																					.13	.87																													
MAD																							.90	.08																											.02
MAT																								.98																											
PEAS																									.85	.13																									
PEACE																									.02	.96																									
SAVE																											.94	.06																							
SAFE																											.24	.74																							
KIT																													.94	.04																					.02
KID																													.06	.90																				.04	
LIST																															1.00																				
TIME																																.94																			.02
SPEED																																	.98	.02																	
DIVE																																	.04	.94																	
DROP																																			.98																
COURSE																																				1.00															
PLOT																																					.98														.02
AIR																																						1.00													
FIRE																																							.98	.02										
LAUNCH																																								1.00										
ATTACK																																									.98									
BINGO																																										1.00								
CANCEL																																											1.00							
DISTANCE																																												1.00						
COPY																																													1.00					
REFUEL																																														1.00				
REPORT																																															.94				.06
PROCEED																																																1.00			
LABEL																																																	1.00		
STATION																																																		.96	.04

45

APPENDIX F

CONFUSION MATRIX FOR OPERATOR MENTAL LOADING – EXPERIMENTAL CONDITION RDO

The matrix element a_{ij} indicates the proportion of the time that the T600 "thought" that a subject said word j when the subject actually said word i. The last column, annotated "beep", indicates the proportion of the time that the T600 rejected vocabulary word i and "beeped".

Several rows do not sum to exactly 1.00; this is because of inaccuracy in the elements of the row caused by rounding to two decimal places.

CONFUSION MATRIX FOR OPERATOR MENTAL LOADING – EXPERIMENTAL CONDITION RD1

The matrix element a_{ij} indicates the proportion of the time that the TS00 "thought" that a subject said word j when the subject actually said word i.) The last column, annotated "beep", indicates the proportion of the time that the TS00 rejected vocabulary word i and "beeped".

Several rows do not sum to exactly 1.00; this is because of inaccuracy in the elements of the row caused by rounding to two decimal places.

	GALE	TALE	CANE	CARE	TIP	DIP	BEAT	PEAT	GOLD	COLD	BARK	PARK	BIG	PIG	TEN	DEN	SAP	SAT	RACE	RAZE	LAKE	LATE	MAD	MAT	PEAS	PEACE	SAVE	SAFE	KIT	KID	LIST	TIME	SPEED	DIVE	DROP	COURSE	PLOT	AIR	FIRE	LAUNCH	ATTACK	SINK	CANCEL	DISTANCE	COPY	REFUEL	REPORT	PROCEED	LABEL	STATION	BEEP*
GALE	.91	.09																																																	
TALE	.10	.88																																																	.04
CANE	.02		.67	.31																																															.02
CARE			.15	.83																																															.02
TIP					.54	.31																																													.02
DIP					.33	.56																																													
BEAT							.74	.17																					.06																						.02
PEAT			.02				.29	.62																					.08																						
GOLD									.90	.08																																									.02
COLD									.28	.70															.04					.02																					
BARK																																																			
PARK											.75	.25																							.02																.02
BIG													.73	.21									.02																												.02
PIG													.21	.72																																					
TEN														.02	.81	.17																																			
DEN															.15	.81																																			.04
SAP																	.85	.15																																	
SAT																	.02	.96																																	
RACE																			.94	.06																															
RAZE																			.07	.93																															
LAKE																				.02	.79	.17																			.02										
LATE																					.06	.92																													
MAD																							.88	.08																											
MAT																								.96																											
PEAS																									.90	.10																									
PEACE																									.13	.87																									
SAVE																											.85	.15																							
SAFE																											.27	.69																							
KIT																													.86	.06																					
KID																													.06	.92																					
LIST																															.98																				
TIME																																.94																			
SPEED																																	.92																		
DIVE																																.02		.98																	
DROP																																			.98																
COURSE																																				1.00															
PLOT																																					.85			.06											
AIR																																						1.00													
FIRE																																							.60	.98											
LAUNCH																																								.98											
ATTACK																																									.86	.00									
BINGO																																										1.00									
CANCEL																																											.98								
DISTANCE																																												.96							
COPY																													.02																	.98					
REFUEL																																														.98					
REPORT																																															.87				.04
PROCEED																																																1.00			
LABEL																																																	1.00		.13
STATION																													.02												.02										.98	

APPENDIX H

CONFUSION MATRIX FOR OPERATOR MENTAL LOADING – EXPERIMENTAL CONDITION RD2

The matrix element a_{ij} indicates the proportion of the time that the Doob "thought" that a subject said word j when the subject actually said word i. The last column, annotated "beep", indicates the proportion of the time that the Doob reported vocabulary word i and "beeped".

Several rows do not sum to exactly 1.00; this is because of inaccuracy in the elements of the row, caused by rounding to two decimal places.

APPENDIX I

CONFUSION MATRIX FOR ALL OPERATOR MENTAL LOADING – EXPERIMENTAL CONDITIONS COMBINED

The matrix element a_{ij} indicates the proportion of the time that the T600 "thought" that a subject said word j when the subject actually said word i.¹ The last column, annotated "beep", indicates the proportion of the time that the T600 rejected vocabulary word i and "beeped".

Several rows do not sum to exactly 1.00; this is because of inaccuracy in the elements of the row caused by rounding to two decimal places.

49

APPENDIX J

SUBJECT VERBAL ERRORS*

An entry w/x (y/z) indicates that a total of w subject verbal errors, of which y were errors of not speaking any word or speaking a non-vocabulary word (when prompted with a vocabulary word), occured in the first half of the trial and a total of x subject verbal errors, of which z were errors of not speaking any word or speaking a non-vocabulary word (when prompted with a vocabulary word), occured in the second half of the trial.

SUBJECT NUMBER	NRT		RD0		RD1		RD2	
1	0/0	(0/0)	0/0	(0/0)	1/0	(0/0)	4/2	(4/2)
2	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)	1/1	(1/1)
3	0/0	(0/0)	1/0	(1/0)	1/2	(1/0)	1/1	(1/1)
4	0/1	(0/0)	0/0	(0/0)	3/3	(2/2)	0/2	(0/0)
5	1/1	(0/0)	0/0	(0/0)	0/0	(0/0)	0/1	(0/0)
6	1/0	(1/0)	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)
7	0/0	(0/0)	0/0	(0/0)	4/2	(2/2)	0/1	(0/1)
8	0/0	(0/0)	0/0	(0/0)	0/1	(0/1)	1/1	(0/1)
9	0/0	(0/0)	0/1	(0/1)	0/1	(0/1)	0/1	(0/0)
10	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)	0/1	(0/1)
11	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)	1/0	(1/0)
12	0/1	(0/1)	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)
13	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)	2/1	(2/0)
14	0/0	(0/0)	0/1	(0/1)	0/0	(0/0)	1/0	(0/0)
15	1/1	(1/1)	2/2	(1/1)	1/1	(1/1)	2/1	(1/1)
16	0/0	(0/0)	2/1	(1/0)	0/0	(0/0)	0/0	(0/0)
17	0/0	(0/0)	1/1	(0/0)	0/0	(0/0)	0/0	(0/0)
18	0/2	(0/0)	2/0	(2/0)	1/1	(1/0)	2/0	(2/0)
19	0/1	(0/0)	0/0	(0/0)	1/0	(1/0)	0/0	(0/0)
20	0/0	(0/0)	0/0	(0/0)	0/1	(0/1)	0/0	(0/0)
21	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)	0/1	(0/1)
22	0/0	(0/0)	0/0	(0/0)	1/0	(1/0)	1/0	(1/0)
23	0/0	(0/0)	1/0	(0/0)	0/0	(0/0)	1/1	(1/0)
24	0/0	(0/0)	0/0	(0/0)	0/0	(0/0)	1/0	(1/0)

OPERATOR MENTAL LOADING - EXPERIMENTAL CONDITION

* A subject verbal error was defined in this research to be a failure of a subject to repeat correctly the presented vocabulary word. This failure could be either a failure to respond (omission) or responding with a non-vocabulary word or the wrong vocabulary word (commission).

Subjects 1 to 16 inclusive had "little experience" on the T600 and subjects 17 to 24 inclusive had "no experience".

APPENDIX K

RATER SCORES

OPERATOR MENTAL LOADING - EXPERIMENTAL CONDITION

SUBJECT NUMBER	RD0	RD1	RD2
1	197	188	90
2	191	174	99
3	198	171	85
4	195	131	102
5	199	182	98
6	186	181	96
7	196	156	98
8	193	171	64
9	188	172	112
10	194	199	144
11	181	188	21
12	188	187	104
13	187	174	71
14	185	172	81
15	192	184	154
16	179	166	-23
17	170	128	73
18	188	191	94
19	150	139	56
20	186	179	91
21	198	182	155
22	195	55	31
23	194	163	110
24	182	170	87

Subjects 1 to 16 inclusive had "little experience" on the T600 and subjects 17 to 24 inclusive had "no experience".

Subjects 22 and 24 each had approximately one half hour prior experience on the RATER; no other subjects had prior experience on the RATER.

To avoid unnecessarily complex instructions, subjects were told that their RATER scores would be simply number of correct responses minus number of incorrect responses, which included both omission and commission errors. This made the RATER tasks more demanding since it discouraged both guessing and failing to respond. However, it is not possible to determine the exact number of errors made from the RATER counters; it is only possible to calculate a lower bound on the number of errors. For this reason, the RATER scores actually assigned were calculated with the following commonly used formula: score = two times number of correct responses minus total number of responses. A perfect score for any experimental condition was 200.

51

APPENDIX L

SUBJECTIVE FATIGUE SCORES*

OPERATOR MENTAL LOADING - EXPERIMENTAL CONDITION

SUBJECT NUMBER	NRT	RD0	RD1	RD2
1	18	18	18	18
2	17	20	17	16
3	13	13	13	13
4	14	19	13	12
5	12	12	14	13
6	15	16	14	14
7	13	13	10	13
8	18	13	10	16
9	10	13	12	12
10	16	13	16	11
11	12	11	12	11
12	21	21	21	21
13	16	16	19	16
14	15	17	13	17
15	11	17	12	18
16	14	12	9	12
17	16	16	16	12
18	16	16	16	16
19	13	12	7	12
20	12	12	12	12
21	16	13	11	12
22	11	11	11	11
23	14	15	14	18
24	12	12	12	9

* Higher scores are associated with lower subjective fatigue and vice versa.

Scores were obtained by multiplying the number of items scored as "better than" by two and adding the number of items scored as "same as", as recommended by those who developed the checklist (Pearson and Byars, 1956).

Subjects 1 to 16 inclusive had "little experience" on the T600 and subjects 17 to 24 inclusive had "no experience".

DISTRIBUTION LIST

No. of Copies

DEFENSE TECHNICAL INFORMATION CENTER   2
CAMERON STATION
ALEXANDRIA, VA   22314

LIBRARY, CODE 0142   2
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA   93940

DEAN OF RESEARCH   1
CODE 012
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA   93940

LIBRARY, CODE 55   1
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA   93940

PROFESSOR GARY POOCK, CODE 55PK   234
OPERATIONS RESEARCH DEPARTMENT
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA   93940